International Journal of Hybrid Information Technology
Vol.8, No.8 (2015), pp.191-200
http://dx.doi.org/10.14257/ijhit.2015.8.8.19

Recommended Study of the Flow of Information based on TF-IDF

Liuqing Li and Rui Zhang

Huanghuai University, Henan, China
Henan Agricultural University, Henan, China
943360673@qq.com

Abstract

The methods to resolve Information overload can mainly classify two kinds: Information Retrieval and Information Filtering. Based on the needs of the people, Information Retrieval will search out the related information and technology from the information stored in some way. Information Filtering will choose out user preferred personalized information from the dynamic information flow based on the filtering requests of the people. The paper analyzed micro-blogs users needs and motivation, According to micro-blogs users motivation, effectively, an approach based on tem frequency inverse document frequency (TF-IDF) was proposed. This article constructed personalized recommendation models in on-line social streams based on ties strength, topic relevance and trust dimensions. The experiments on the Sina blogs data showed that the proposed method could reduce the ranks of irrelevant tweets effectively and achieve better performance than several baseline methods based on cosine and hash tags

Keywords: *Information filtering; Personalized; TF-IDF; Micro-blogs*

1. Introduction

Micro-blog as a network platform of information sharing and communication, has been widely used in recent years. How to let users get micro-blog content of interest in massive micro-blog has become the new research directions of a micro-blogging platform. Most of the current recommendation system for a number of micro-blog users to recommend , and for the micro-blog platform , because of the shorter length of the text micro-blog, diverse interests of user, therefore the effect of recommend is not ideal for the user [1]. Micro-blog lets people obtain real-time and vast amounts of information through a virtual network. Now popular micro-blog website, Twitter abroad, domestic such as sina micro-blog *etc.* There are many study on the micro-blog, such as finding the influence of the highest users in micro-blog; advertise propaganda and investment in the micro-blog; authenticity test micro-blog information, so as to prevent the spread of rumors; and classification of micro-blog information.

Unlike previous social recommendation always use social networking features to recommend new content or links, information retrieval model presented in this paper is recommended for users micro-blog information to meet their interests and preferences and shows a new homepage [2]. Users can find other related micro-blog information, and they can also discover new related user in this homepage.

2. Problem and Research Framework

Social media not only faces the problem of information overload, also faces challenges of information hiding. Social search engines can provide a simple information retrieval approach, however, these engines cannot predict what users like and information of

ISSN: 1738-9968 IJHIT

interest to the exhibition. In order to solve this problem, social search engine and personalized recommendation algorithm combining is necessary.

Personalized micro-blog recommendation is to estimate how much a micro-blog interest value for each user, so how do you handle user's interest is one of the key issues in this article. Users are mainly 3 actions on micro-blog: follow other users, publish micro-blog and forwarded to other users of micro-blog. The user's interests can be found by analyzing the user's behavior, while micro-Bo of user is the most direct reflection of users ' interests, it is also the focus of this paper [3].

3. Commendation based on Information Content Characteristic

Commendation based on information content characteristic is a method which is often used by full text commendation system as well as socialized net information stream commendation. The most important aspect of it is Topic Relevance calculating. When calculating topic relevance, firstly, we should establish characteristic vector of user interest model, that is, to any of the user u, his interest vector can be shown as the formula:

$$V_u = \{v_u(w_1), v_u(w_2), .v_u(w_i)\} \tag{1}$$

i —the extractive topic number of the information in all the microblog users released

$v_u(w_i)$ —the interest degree of user u to the topic w_i

The expression $v_u(w_i)$ can be calculated by the arithmetic TF-IDF:

$$v_u(w_i) = TF_u(w_i) \square IDF_u(w_i) \tag{2}$$

In the formula, the expression $TF_u(w_i)$ shows the frequency of the key word w_i in information user u released.

$$IDF_u(w_i) = \log\left(\frac{U}{u}\right) \tag{3}$$

U —representing the total number of all the users

u —representing the number of the users who have used the key word w_i at least once

Similarly, we can establish topic vector of given information. What the difference is TF have been expressed to show the frequency of certain key word mentioned in all microblog information. At last, we can use cosine similarity arithmetic to calculate the similarity between user interest vector and topic vector of information content.

4. Information Retrieval Model

In the information retrieval model for arbitrary user u and micro-blog $t(t \in T)$, the values $I_u(t)$ that are used to gauge user u interest in micro-blog t, user's interests can be drawn from the historical micro-blog information.

Topics related measure has been widely used in the recommendation system, the flow of information in social networks have begun to successfully use the sort field , and superior to the classical metrics topic model LDA based on the type of text processing micro-blog according to research topics $TF - IDF$ based on the correlation metrics [4]. In

order to calculate the user's topic and a micro-blog information flow correlation vector, first get a topic from the user 's micro-blog post , and then build a similar vector reaching the user 's micro-blog information flow, the final score is calculated on correlations choice the most suitable micro-blog recommended to the user. This part of the collection R_u published by the similarity, and calculate the user's micro-blog and micro-blog flow of information in any subset of T to estimate the value $I_u(t)$. First, it should be given a random sequence micro-blog X, and define a bag of words vector $BT(X) = w_1, w_2, \cdots, w_n$, the words W_i should be contained in at least one t in micro-blog. However, a micro-blog information flow usually contain a small amount of micro-blog and each micro-blog are less than 140 words, the bag of words vector set up in this way is very sparse, and the similarity cannot be compared. Based on the bag of words in the vector by merging some words make the bag of words Vector final richer. Thus , the model also defines a sequence of words vector $BP(X) = P_1, P_2, \cdots, P_n$, at least ,sequence p_i pi words appear in the same micro-blog t. For a micro-blog sequence X in the case of words and word sequences given two scoring formulas.

Definition 4 is given a word w and a micro-blog information flow $Xscore(w, X)$ represent w value in a given X.

$$\begin{cases} tscore(w, X) = TF_T(w, X) \cdot IDF_T(w) \\ IDF_T(w) = \log \dfrac{|T|}{DF_T(w)} \end{cases} \tag{4}$$

$TF_T(w, X)$ is micro-blog number, which contains the words w in micro-blog sequence X, and DF_T is micro-blog number, which contains the words w in micro-blog information flow T. $TF - IDF$ is a statistical method used to evaluate a set of words for a file or a corpus in which the importance of a document. $TF - IDF$ can be seen in the experimental section in dealing with some long blog, which has more advantages. Definition 5 given a set p and a micro-blog X, $pscore(p, X)$ represent the word p in a given x.

$$\begin{cases} pscore(p, X) = TF_T(p) \cdot IDF_P(p) \\ IDF_p = \log \dfrac{|T|}{DF_P(p)} \end{cases} \tag{5}$$

TF is the number of micro-blog in information micro-blog flow, which contains word sequences of p, $DF_P(p)$ is the number of micro-blog in information micro-blog flow T, which contains word sequences of p. Micro-blog is usually very short document which contains a few sentences, but a few words or a combination of the words appear at the same time, there is a high probability in two micro- Bo. Finally we combine the definitions 4 and definitions 5 to get the score formula of estimated two Micro-blogs sequence similarity.

Definition 6 give two micro-blog information flow X_1 and X_2, we define the similarity as $\text{int}_A(X_1, X_2)$.

$$\begin{aligned} \text{int}_A(X_1, X_2) = (1-\lambda) \sum_{w \in BT(X_1) \cap BT(X_2)} tscore(w, X_2) \\ \lambda \sum_{p \in BP(X1) \cap BP(X_2)} pscore(p, X2) \end{aligned} \tag{6}$$

In the calculation of similarity, if $\lambda = 0$ then only consider a single word, on the contrary, if $\lambda = 1$ only consider the situation of the words. In the calculation of int_λ,

which can use different λ values to determine the optimal scheme. After a given user u and the published micro-blog set R_u, for any value of $(\lambda_1, \lambda_2, \cdots, \lambda_i)\lambda \in [0,1]$, the definition of $I_u(X) = \text{int}_\lambda(X, R_u)$,which can estimate the interest degree of user u for micro-blog sequence X. So that this can be calculated the value int_λ in the maximum for R_u to choose the micro-blog information flow T, and micro-blog k is recommended to the user.

5. The Cold Start Problems

Information retrieval model presented in this paper is the main user of the information published micro-blog R_u, so that you can build up a rich language models. There are a considerable number of inactive users micro-blog users, they gave very little or no published micro-blog. Interest-based collaborative filtering can draw the user's thinking , not only in the field to get the user's own , you can also get through the user's social relationships, that is, the user concern micro-blog friends to get .In fact ,unlike other social networks, micro-blogging focus on social relationships based on interest in sharing built. In building a user model some of my friends are more important than others, because they provide more important and relevant information. The definition 7 of an algorithm has been improved

$$tscore\ (w, X) = IDF_T(w) \cdot \sum_{x \in X} auth\ (ux) \cdot in(w, x) \tag{7}$$

If micro-blog X contains the word w, then the value of $in(w, x)$ is 1 , otherwise 0, $auth(u_k)$ refers to originally published micro-blog w of the influence of user u, this part will be discussed in the next section . The same definition of 2 modified and then consider a group of words.

User 's influence on the impact of micro-blog users has a clear definition ,which is a very difficult thing. Users are usually influential producers of information, often publish some of the other interesting or important users of micro-blog. Such users often use the number of its followers and those who are concerned about the measure. Although there are some other good measure of user characteristics, such as the number of influential micro-blog users forwarding , etc., from the perspective of simple and achieve, this paper no longer consider other features. For each user u, we calculated the $auth(u) \in [0,1]$ values through the influence of a linear formula.

$$auth\ (u) = A \cdot \frac{1}{1 + e^{\frac{ffRatio\ (u)}{\alpha}}} + (1 - A) \cdot \frac{1}{1 + e^{\frac{foll\ (u)}{\beta}}} \tag{8}$$

Among them: $ffRatio\ (u) = foll\ (u) / fri(u)$, $foll\ (u)$ and $fri(u)$ respectively means the concerned number and the concern number of the user's u, and $A \in [0,1]$, α and β is the logical parameters. Values for all the parameters in an artificial influence on user testing is completed on the sort of small data sets. Tested: $A = 0.5$, $\alpha = 2$, $\beta = 2000$.

6. Micro-blog Recommendation Algorithm based on TF-IDF

Through the above discussion, here are based on the basic idea of TF-IDF micro-blog recommendation algorithm:

First of all micro-blog information flow T received from the user in selecting any k micro-blog, which consist of a micro-blog sequence S, and get the word w_i on micro-blog

sequence S word, some words which frequently occur together are combined to generate word sequences p_i.

Through the calculation of $tscore(w, S)$ and $pscore(p, S)$ (such as equation (2) and (3) below) can get the words w_i or word sequences p_i on micro-blog sequence S in the information flow in T ,which marked degree and get the theme of micro-blog sequence S.

After getting micro-blog theme of the sequence s by calculating the user's interest most k posts sequence s (s $\in$ t), which is recommend to the user (such as equation (1) below).

Interest in the idea of collaborative filtering algorithm based on the assumption that the user can get past its micro-blog published in a collection of R_u , and the similarity is defined as a sequence of two micro-blog $\text{int}_\lambda(X_1, X_2)$,so if $X_1 = S, X_2 = R_u$,for any $\lambda \in [0,1]$, defined $I_u(X) = \text{int}_\lambda(X, R_u)$ to estimate the user u towards micro-blog sequence S of degree of interest , and then by calculating the maximum value int_λ , in terms of R_u to elect k micro-blog of micro-blog information flow in T recommended to the user.

Finally, the model takes into account the user's cold start problems and influence, and the algorithm was improved, the last algorithm $\text{int}_\lambda + auth(u)$ is used to study the influence of users of micro-blog information recommended by the impact.

Personalized commendation model of micro-blog stream based on users' motivation.

Given this, the study establish a systematic an comprehensive micro-blog information stream commendation model in term of three dimensions-topic relevance, strength of ties and credibility, expected to improving the shortcomings of single commendation model when the users are in the state of various of cognitive motivations, as shown in Figure 1.

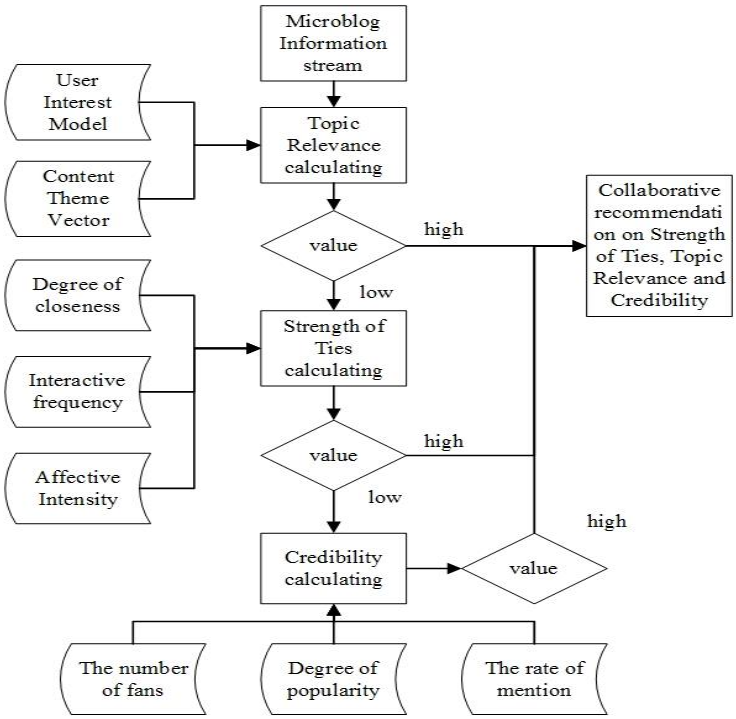

Figure 1. Personalized Commendation Model of Micro-Blog Stream based on Users' Motivation

Algorithm design

Input: user access to key words each time the user query log in;

Output: the similarity with the query keywords vector existing values in the database is not paper and similarity of 0, and according to the similarity value from big to small order;

1) extracted from each Webpage keywords as the feature word, and these feature words and keywords query every time the user binding, rearrangement and according to a lexicographic order, combined together to form a standard feature set of words;

For example, there are Webpage document set (NDoc1, NDoc2, NDoc3, NDoc4), all the feature words together for (W1, W2, W3, W4, W5, W6), and when the query words (WQ1, WQ2), where wq1= W4, the standard set of words as features (W1, W2, W3, W4, W5, W6, WQ2), Webpage document feature item vocabulary matrix in Table 1.

Table 1. Webpage Document Vector Space Model

	w1	W2	W3	W4	W5	W6	wq2
Q1	0	0	0	0	0	0	1
NDoc1	14	21	33	0	0	0	0
NDoc2	0	11	15	0	0	22	0
NDoc3	8	0	0	14	15	17	0
NDoc4	0	8	9	12	0	15	0

2) t_j is calculated for each term tf_{ij} appears in the Webpage text d_i frequency; calculation of $\log(N / n_k + 0.5)$ words formula of inverse document frequency idf_i ; then the formula with weight (3.1) to calculate the weight of each feature words each Webpage of document vectors, forming the Webpage document vector in a vector space;

3) to calculate the similarity of each document vector and the query vector Webpage between the cosine coefficient method, see equation (2). The interception of similarity values greater than 0.2000 articles, and from high to low return results;

Association rule mining is used to find correlation between the attributes of databases. Association rules is the initial motive of shopping basket analysis problem, the goal is to find the different commodities of association rules mining in transaction database, the relationship between the useful knowledge description data item value [5-7]. These knowledge characterizes customer buying behavior and mode, use these rules, can effectively guide the scientific arrangement and design business purchase goods shelves. The form of association rule is a rule is, "to buy milk and bread customers, 90% of people bought butter", namely "(milk, bread)→ butter" issue.

Let be the $I = \{i_1, i_2, ..., i_m\}$ set of items. A related task data D is a collection of database transactions, where each transaction is a set of T $T \subseteq I$, so. Each transaction is an identifier, called TID. Let A be a set of transaction, $A \subseteq T$ T contains A if and only if. Association rules are shaped implication $A \Rightarrow B$, such as one $A \subset I, B \subset I$, and $A \cap B = \varnothing$. Rule $A \Rightarrow B$ D in the transaction set, with the support of S, where s is the D transaction contains the percentage of A ∪ B, namely P (A, B). Rule $A \Rightarrow B$ D C has confidence in the transaction set, where C is contained in the D A transaction also includes a percentage of B, namely P (B |A).

$$Support(A \Rightarrow B) = P(A \cup B) \quad (9)$$

$$Confidence(A \Rightarrow B) = P(B \mid A) \quad (10)$$

The support and confidence are two important concept description of association rules, the former for statistical measure of the importance of association rules in the data, said the rules which is used to measure frequency; credible degree of association rules, said the strength of the rules. In general, only the support and confidence of association rules are high may be only the interesting rules, useful. Association rules mining is mainly realized by the two steps:

Step 1, according to the minimum support degree to find the database in D all the frequent item sets.

Step 2, according to the frequent item sets and minimum confidence generated Association rules.

Task one step is to quickly and efficiently find all frequent item sets in D, is the central problem of the association rule mining algorithm of association rules mining, is a measure of the standard; step two, relatively easy to achieve, so now all association rules mining algorithm is designed for the first step forward.

7. Experimental Results and Analysis Data

The experimental data are collected through a self- extracting and Sina micro-blog open API. The data set includes 180,000 posts information. After cleaning up the data set after removal of junk information used in this experiment [8-9].First it should clean out micro-blog of fewer than 10 characters, then get rid of less than three nouns micro-blog , and finally remove to begin with @ micro-blog information, because these micro-blog user's private conversations are actually unsuitable recommendations to other users.

8. Algorithm Evaluation and Methods Comparison

This section through experimentally compared intλ and based on cosine similarity and label vector algorithm performance and evaluation mechanism using a variety of automatic evaluation of the intλ effect. Experiment randomly select 200 users from a centralized data , for each user, select its 90 % of the published micro-blog as a measure of the user's own interest (thesis for R_u) data , and the remaining 10% of the part and the other all users of the micro-blog mixed together to form a paper test data sets. Assuming paper recommendation system will test, if the data set 10% of the user's own micro-blog recommended to the user is called the right recommendation.

This paper uses a standard evaluation methods for information retrieval as a mechanism for the evaluation of information retrieval models. We calculate for each user: P@k, precision arithmetic refers to the proportion of pre- k micro-blog of right recommended;

S@k, the success rate is at least in the first k posts, there is a correct proportion of micro-blog;

MRR, average ranking algorithm is to last in the final sorted sequence results in reciprocal Q, and each of the article the correct location of the micro-blog the average (if (11) shown above), high MRR indicating a higher accuracy of the algorithm.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (11)$$

Firstly, in order to calculate λ value optimization in intλ, this experiment from the data set 200 users using the above parameters are chosen for testing. Figure 2 P@k, Figure 3 MRR and Figure 4 S@k under different values of λ is displayed the performance results $_{int\ \lambda}$.

Results show that when λ values is close to 1, pscore has more weight, and corresponding to the model in the evaluation of algorithms is also better. In fact, the maximum value P @ k and S @ k is obtained when the λ = 1. In Figure 3, when λ =0.9, in the values of λ, MRR shows a downward trend, which corresponds to the correct micro-blog's ranking is on the rise. Therefore, when λ = 0.9, $_{int\ \lambda}$ on the accuracy and success rates are doing well and will be interesting micro-blog arranged in a higher position.

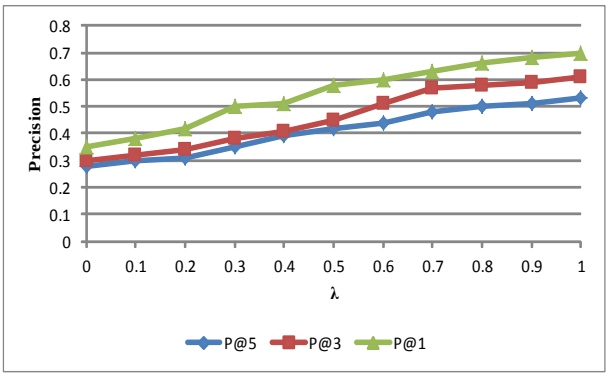

Figure 2. P@ k Values with Different λ in Information Model

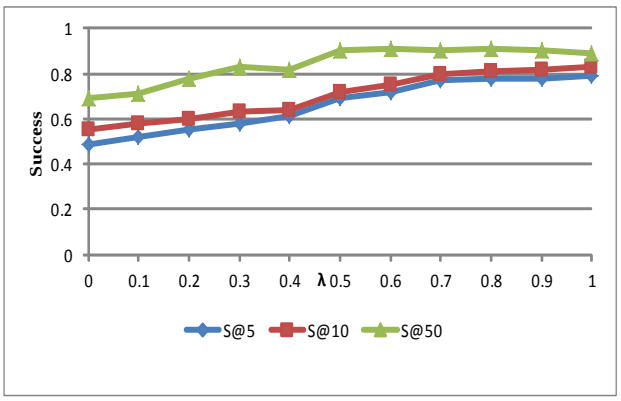

Figure 3. S@ k Values with Different λ in Information Model

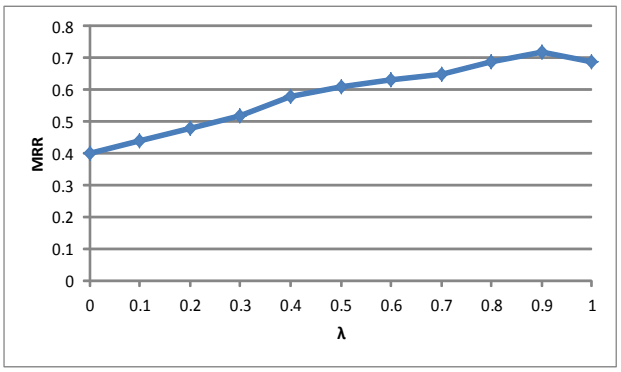

Figure 4. MRR Values with Different λ in Information Model

Then compares the int0.9 method and the existing cosine similarity and label vectors. Cosine similarity is used to measure the similarity between two vectors by measuring the angle between the inner product space of the cosine value. The interest degree of each micro-blog in micro-blog information flow T in this paper is determined by calculating t and T_u cosine similarity of word frequency vectors derived. Method and cosine vector similarity method is similar to the label, not just vectors consisting of words but is composed of micro-blog user label. This method is similar to the literature F. Abel and other proposed method is used to label a concise summary of the topic of micro-blog words, usually by the user to add their own.

As can be seen from table 1, int0.9 performs best in a variety of evaluation mechanisms. Especially in the P @ 1 metric, int0.9 would rank a correct micro-blog in the first position on the accuracy rate of 69%. In addition int0.9 MRR measure reached 73% accuracy rate, which shows the correct method to use int0.9 micro-blog ranked higher than average position than the other two methods.

Finally, the algorithm will make int0.9 methods and user influence combined by calculating int0.9 + auth (u) to explore the influence of users on the micro-blog recommended by the affect the flow of information . It was evaluated by the MRR and int0.9 + auth int0.9 effect (u) of the two methods, the results shown in the MRR int0.9 + auth value (u) Method to achieve higher int0.9 MRR value under 0.92. Thus, the user's influential friends has issued micro-blog, which is the key of the micro-blog information recommendation.

Table 1. Comparison of Methods of int0.9, Cosine Similarity and Label

	P@1	P@3	P@5	MRR
Int0.9	0.69	0.60	0.53	0.73
Cos	0.30	0.26	0.22	0.46
tag	0.33	0.28	0.21	0.51

9. Conclusion

This paper presents a method to measure user's interest in micro-blog, and combine collaborative filtering for micro-blog users to recommend more interesting kind of micro-blog information, and a recommendation system can be applied to many different scenarios, which provides users with more interesting home pages, and, experimental results show that the recommended method in this paper has a high accuracy rate. In the future, we will consider the user or micro-blog features to further improve the precision of micro-blog recommended. In the future, we will further study the relationship between micro-blog users and micro-blog cold start System.

Acknowledgement

The project is supported by Economic management experimental teaching demonstration center in Huanghuai University.

References

[1] F. Sebastian, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, vol. 34, no. 1, **(2002)**, pp. 24-47.

[2] H. Zhuang, "The Knowledge grid", World Scientific, Singapore, **(2004)**.
[3] Y. Jia, "Semantic Link Network Builder and Intelligent Browser", Concurrency and Computation: Practice and Experience, vol. 16, no. 14, **(2004)**, pp. 1453-1476.
[4] J. Chen and E. Chi, "Speak little and well: Recommending Conversations in Online Social Streams", Proceedings of the 2011 annual Conference on Human Factors in Computing Systems, ACM, **(2011)**, pp. 217-226.
[5] A. N. Joinson, "Looking at, Looking up or Keeping up With People: Motives and Use of Facebook", Proceedings of the SIG-CHI conference on Human Factors in Computing Systems, ACM, **(2008)**, pp. 1027-1036.
[6] P. Mika and G. Tununarello, "Web Semantics in the Intelligent Systems", IEEE, vol. 23, **(2008)**, pp. 82-87.
[7] C. Hewitt, "ORGs for Sealable, Robust, Privacy-Friendly Client Internet Computing", IEEE, **(2008)** September-October, pp. 96-99.
[8] T. C. Jepson, "The basics of reliable distributed storage networks IT Professional", vol. 6, **(2004)** May-June, pp. 18-24.
[9] M. W. Beny, T. Do, GW.O, Brien, V. Krishna and S. Varadhan, "SVDPACKC User, 5 Guide, University of Tennessee, **(1993)** April.
[10]

Authors

Liu-qing Li, born in June, 1981, Biyang, Henan, P R China
Current position: University Teachers;
grades: lecturer:
Scientific interest: Computer science and technology:
Scientific interest: Computer application
Publications: Research of Blind Mixed Signal Separation Technology Based on Fixed Step Size Natural Gradient Algorithm (Volume12, Number15, 2013 INFORMATION TECHNOLOGY JOURNA) etc. I've completed 12 theses since 2009, Include 10 core paper.
Experience:
1999.09 - 2003.07 Henan University of Finance and Economics information management and information system of professional;
2003.07 - 2014.09 Huanghuai University engaged in teaching and research work;
2005.09 - 2008.07 Huazhong University of Science and Technology for master's degree;
2003.07-2014.11, more than 20 papers published, presided over the participation of 8 scientific research items, presided over the participation of 4 utility model patents, jointly compiles 4 teaching materials

Rui Zhang
Current position, grades: Lecturer,
University studies: Studying Computer Science at Henan Agricultural University (China)
Scientific interest: Computer Multimedia Technology

